HYPER-PARAMETER SELECTION IN BAYESIAN STRUCTURAL EQUATION MODELS

Hirose, Kei
Graduate School of Mathematics, Kyushu University

Kawano, Shuichi
Institute of Medical Science, University of Tokyo

Miike, Daisuke
Graduate School of Mathematics, Kyushu University

Konishi, Sadanori
Department of Mathematics, Chuo University

https://doi.org/10.5109/25906

出版情報：Bulletin of informatics and cybernetics. 42, pp.55-70, 2010-12. Research Association of Statistical Sciences
バージョン：
権利関係：

HYPER-PARAMETER SELECTION IN BAYESIAN STRUCTURAL EQUATION MODELS

by

Kei Hirose, **Shuichi Kawano**, **Daisuke Miike**
and
Sadanori Konishi

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.42

FUKUOKA, JAPAN
2010

Bulletin of Informatics and Cybernetics, Vol. 42, 2010

HYPER-PARAMETER SELECTION IN BAYESIAN STRUCTURAL EQUATION MODELS

By

Kei Hirose[*], **Shuichi Kawano**[†], **Daisuke Miike**[‡]
and
Sadanori Konishi[§]

Abstract

In the structural equation models, the maximum likelihood estimates of error variances can often turn out to be zero or negative. In order to overcome this problem, we take a Bayesian approach by specifying a prior distribution for variances of error variables. Crucial issues in this modeling procedure include the selection of hyper-parameters in the prior distribution. Choosing these parameters can be viewed as a model selection and evaluation problem. We derive a model selection criterion for evaluating a Bayesian structural equation model. Monte Carlo simulations are conducted to investigate the effectiveness of the proposed modeling procedure. A real data example is also given to illustrate our procedure.

Key Words and Phrases: Bayesian approach, Improper solutions, Model selection criterion, Prior distribution, Structural equation modeling

1. Introduction

Structural equation models that include the factor analysis model and model in path analysis play an essential role in various fields of research such as social, educational, behavioral and biological sciences, public health, and medical research (see, e.g., Bentler and Stein, 1992; Jörekog and Sörbom, 1996; Pugesek *et al.*, 2003; Xiong *et al.*, 2004; Liu *et al.*, 2008).

The structural equation model is usually estimated by maximum likelihood methods under the assumption that the observations are normally distributed. In practice, however, the maximum likelihood estimates of error variances can often turn out to be zero or negative. Such estimates are known as improper solutions, and many authors have studied these inappropriate estimates both from a theoretical point of view and

[*] Graduate School of Mathematics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan. k-hirose@math.kyushu-u.ac.jp. Research Fellow of the Japan Society for the Promotion of Science.

[†] Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan. skawano@ims.u-tokyo.ac.jp. Research Fellow of the Japan Society for the Promotion of Science.

[‡] Graduate School of Mathematics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan. Present address: Soft service Co., Ltd. 3-3-22, Hakataeki-Higashi, Hakata-ku, Fukuoka 812-0013, Japan.

[§] Department of Mathematics, Chuo University, 1-3-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan. konishi@math.chuo-u.ac.jp

also by means of numerical examples (see, e.g., van Driel, 1978; Anderson and Gerbing, 1984; Boomsma, 1985; Gerbing and Anderson, 1987; Chen *et al.*, 2001; Flora and Curran, 2004). In order to prevent the occurrence of improper solutions in structural equation model, we employ a Bayesian approach by specifying a prior distribution for error variances.

An essential point in the Bayesian approach is the choice of a prior distribution. In the factor analysis model which is the special case of structural equation model, some prior distributions have been proposed by earlier authors (see, e.g., Martin and McDonald, 1975; Akaike, 1987; Hirose *et al.*, 2010; Yoshida and West, 2010). Akaike (1987) introduced a spherical normal distribution of the standardized factor loadings, which is theoretically derived from the information extracted from the knowledge of the likelihood function, and numerical examples show that it can prevent the occurrence of improper solutions. Thus, Akaike's (1987) prior distribution seems to be attractive. It is, however, difficult to apply his prior distribution directly to the structural equation models, since the covariance structure of structural equation model is too complex to derive the standardized spherical normal distribution. Hirose *et al.* (2010) considered a prior distribution for unique variances in factor analysis model and used exponential distributions for the inverses of unique variances. In this paper we derive an inverse exponential distribution for error variances in the structural equation models according to the basic idea given by Akaike (1987) and Hirose *et al.* (2010). The model is then estimated by posterior modes.

In the Bayesian structural equation models, the hyper-parameters in the prior distribution are often subjectively given. However, the modeling procedure based on such subjective hyper-parameters does not always provide appropriate estimates. Therefore, it is important to select suitable values of hyper-parameters by using an information extracted from the data. Choosing these parameters can be viewed as a model selection and evaluation problem. The AIC (Akaike, 1987), BIC (Schwarz, 1978) and other selection criteria (e.g., Bozdogan, 1987; Ninomiya *et al.*, 2008) cannot be directly applied to the Bayesian structural equation model since these criteria cover only models estimated by the maximum likelihood methods. In this paper, we derive a model selection criterion from a Bayesian point of view (Konishi *et al.*, 2004) for evaluating Bayesian structural equation models. The proposed modeling procedure is investigated by conducting Monte Carlo simulations and analyzing a real data. Numerical results show that our modeling strategy prevents the occurrence of improper solutions and often yields stable estimates.

The remainder of this paper is organized as follows: Section 2 describes maximum likelihood methods for structural equation model. In Section 3, we introduce a Bayesian structural equation modeling. Section 4 describes Monte Carlo simulations to investigate the performance of our modeling procedure. Section 5 illustrates the proposed procedure with a real data example. Some concluding remarks are given in Section 6.

2. Maximum likelihood procedure for structural equation model

A number of models for the analysis of covariance structure, such as LISREL (Bock and Bargmann, 1966; Jöreskog, 1970), EQS (Bentler and Weeks, 1980) and RAM (McArdle, 1980; McArdle and McDonald, 1984), have been proposed. We use the RAM model because the description of this model is quite simple and it generalizes the LISREL and EQS.

First, we define p-dimensional observable random vector, m-dimensional latent vari-

ables and p-dimensional error variables given in the following:

$\boldsymbol{f} = (f_1, \cdots, f_m)'$: m-dimensional latent random vector,
$\boldsymbol{x} = (x_1, \cdots, x_p)'$: p-dimensional observable random vector,
$\boldsymbol{t} = (\boldsymbol{f}', \boldsymbol{x}')'$: $q\ (= m+p)$-dimensional structural variables that include latent variables and observable variables, and they satisfy $\mathrm{E}[\boldsymbol{t}] = \boldsymbol{0}$,
$\boldsymbol{d} = (d_1, \cdots, d_m)'$: m-dimensional error variables for latent variables $\boldsymbol{f}$,
$\boldsymbol{e} = (e_1, \cdots, e_p)'$: p-dimensional error variables for observable variables $\boldsymbol{x}$,
$\boldsymbol{u} = (\boldsymbol{d}', \boldsymbol{e}')'$: q-dimensional error variables with $\mathrm{E}[\boldsymbol{u}] = \boldsymbol{0}$ and $\mathrm{cov}[\boldsymbol{u}] = \Sigma_{\boldsymbol{u}}$. Assume that unknown error variances in $\Sigma_{\boldsymbol{u}}$ are $\sigma_1^u, \cdots, \sigma_v^u$.

The structure between latent variables and observable variables in the RAM model is given by

$$\boldsymbol{t} = A\boldsymbol{t} + \boldsymbol{u}, \tag{1}$$

where $A = (a_{ij})$ is a $q \times q$-coefficient matrix for structural variables. Note that the diagonal elements of A are zeros because the path from a variable to the same variable does not make any sense.

Next, we calculate the variance-covariance matrix of $\boldsymbol{x}$. Suppose that there exists an inverse matrix $T = (I_q - A)^{-1}$, where I_q is a $q \times q$ identity matrix. From Equation (1), we have

$$\boldsymbol{t} = T\boldsymbol{u}.$$

The observable random vector $\boldsymbol{x}$ is then given by

$$\boldsymbol{x} = GT\boldsymbol{u},$$

where G is a $p \times q$-matrix which extracts the observable variables from the structural variables: $G = [\boldsymbol{O}_{p\times m} \quad I_p]$, with $\boldsymbol{O}_{p\times m}$ being $p \times m$ 0-matrix. Then, the variance-covariance matrix of $\boldsymbol{x}$ is given by

$$\Sigma(\boldsymbol{\theta}) = GT\Sigma_{\boldsymbol{u}}T'G',$$

where $\boldsymbol{\theta}$ is a k-dimensional unknown parameter vector. The unknown parameters in the structural equation models are the coefficient matrix A and a lower triangular part of variance-covariance matrix $\Sigma_{\boldsymbol{u}}$. Note that most of the elements of A and $\Sigma_{\boldsymbol{u}}$ are fixed by 0 or 1 according to researcher's hypothesis. The parameter vector $\boldsymbol{\theta}$ is constructed by eliminating these fixed parts.

The structural equation model is usually estimated by the maximum likelihood procedure. Suppose that we have a random sample of N observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ from the p-dimensional normal population $N_p(\boldsymbol{0}, GT\Sigma_{\boldsymbol{u}}T'G')$. The log-likelihood function is then given by

$$\log f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N|\boldsymbol{\theta}) = -\frac{N}{2}\left\{p\log(2\pi) + \log|\Sigma(\boldsymbol{\theta})| + \mathrm{tr}(\Sigma(\boldsymbol{\theta})^{-1}S)\right\}, \tag{2}$$

where S is a sample variance-covariance matrix

$$S = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n'.$$

The maximum likelihood estimates of $\boldsymbol{\theta}$ are given as the solutions of

$$\frac{\partial \log f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

Since the solutions cannot be expressed in a closed form, the quasi-Newton's method is usually used to obtain the maximum likelihood estimates.

In practice, however, the maximum likelihood estimates of error variances can often turn out to be zero or negative, which have been called improper solutions. In order to overcome this difficulty, we take a Bayesian approach by specifying a prior distribution for error variances.

3. Bayesian structural equation modeling

In this section, we investigate the prior distribution for the variances of error variables, and then illustrate a selection procedure of the hyper-parameters in the prior distribution.

3.1. Prior distributions

An important point in the Bayesian structural equation models is the selection of a prior distribution. In the factor analysis model, Hirose *et al.* (2010) derived exponential distributions for the inverses of unique variances according to the basic idea given by Akaike (1987), and numerical examples showed that their prior distributions can prevent the occurrence of improper solutions. On the basis of their prior distributions, we use an inverse exponential distribution for error variances given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \prod_{i=1}^{v} \frac{N\lambda_i}{(\sigma_i^u)^2} \exp\left(-\frac{N\lambda_i}{\sigma_i^u}\right) \qquad (\sigma_i^u > 0 \ \text{ for } \ i = 1, \cdots, v), \tag{3}$$

where $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_v)'$ is a v-dimensional hyper-parameter vector with $\lambda_i > 0$ ($i = 1, \cdots, v$). Note that this prior distribution is an inverse gamma prior distribution

$$\pi(\sigma_i^u|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma_i^u)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma_i^u}\right)$$

with $\alpha = 1$ and $\beta = N\lambda_i$, where $\Gamma(\cdot)$ is a gamma function. Figure 1 shows the inverse exponential distribution when $N = 100$ and $\lambda_i = 0.005$. It can be seen from Figure 1 that a probability that σ_i^u is close to 0 is extremely low. Thus, this prior distribution may prevent the occurrence of improper solutions.

The posterior distribution based on the prior distribution in (3) is

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N; \boldsymbol{\lambda}) &= \frac{f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\displaystyle\int f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}) d\boldsymbol{\theta}} \\ &\propto f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}).\end{aligned}$$

In this paper, the parameters $\boldsymbol{\theta}$ are estimated through modes of the posterior distribution. The procedure is equivalent to obtain estimates by maximizing the penalized log-likelihood function

$$\ell_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \log f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N|\boldsymbol{\theta}) - H_{N,\boldsymbol{\lambda}}(\boldsymbol{\theta}), \tag{4}$$

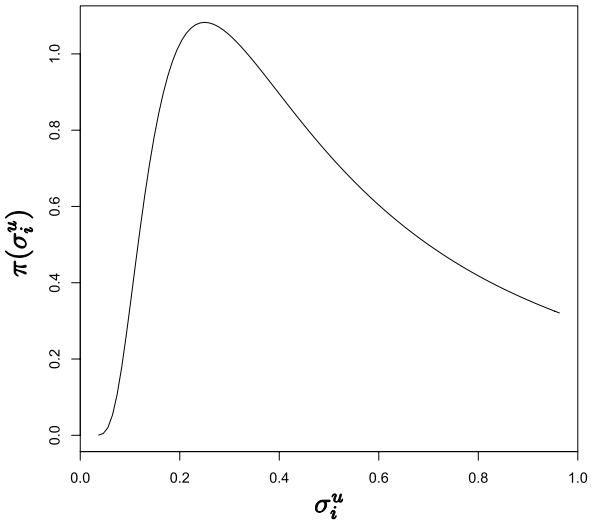

Figure 1: The inverse exponential distribution when $N = 100$ and $\lambda_i = 0.005$.

where $H_{N,\boldsymbol{\lambda}}(\boldsymbol{\theta})$ is a penalty term given by the following:

$$H_{N,\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \sum_{i=1}^{v} \left(\frac{N\lambda_i}{\sigma_i^u} + 2\log\sigma_i^u \right),$$

and the hyper-parameters $\boldsymbol{\lambda}$ can be considered as regularization parameters. Since it is difficult to obtain the parameters that maximize the function in (4) analytically, we use a quasi-Newton's method to obtain the maximum penalized likelihood estimates.

3.2. Model selection criterion

This subsection describes a selection process of hyper-parameters in the prior distribution. For example, the maximum likelihood estimate of σ_1^u becomes negative. When the value of λ_1 is very small, the penalized maximum likelihood estimate of σ_1^u may be close to 0. On the other hand, when the value of λ_1 is large, the penalized maximum likelihood estimate of σ_1^u also becomes large. Therefore, a crucial aspect of model construction is the choice of the regularization parameter $\lambda_1, \cdots, \lambda_v$. In this paper, we derive a model selection criterion GBIC (Konishi *et al.*, 2004) for evaluating Bayesian structural equation models. The proposed procedure enables us to choose adjusted hyper-parameters $\lambda_1, \cdots, \lambda_v$. For model selection criteria we refer to Konishi and Kitagawa (2008) and references given therein.

The model selection criterion GBIC for evaluating the Bayesian structural equation

model is given by

$$\begin{aligned}\mathrm{GBIC} = &-k\log(2\pi) + k\log N + \log|J_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}})| + N\left\{p\log(2\pi) + \log|\Sigma(\hat{\boldsymbol{\theta}})| + \mathrm{tr}(\Sigma(\hat{\boldsymbol{\theta}})^{-1}S)\right\} \\ &-2\sum_{i=1}^{v}\log\left\{\frac{N\lambda_i}{(\hat{\sigma}_i^u)^2}\right\} + 2\sum_{i=1}^{v}\frac{N\lambda_i}{\hat{\sigma}_i^u}, \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}_i^u$ $(i = 1, \cdots, v)$ are the posterior modes, and $J_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}})$ is $k \times k$ matrix

$$J_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}) = -\frac{1}{N}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left\{\log f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N|\boldsymbol{\theta}) + \log\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})\right\}\Big|_{\hat{\boldsymbol{\theta}}}\right]. \tag{6}$$

The derivation of $J_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ is given by (A6) in Appendix A.

When we have several candidates for hyper-parameter vectors $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_v)'$, the GBIC is calculated for each candidate, and then the model which minimizes the value of GBIC is selected. However, if the dimension of hyper-parameters v is large, it is difficult to calculate the GBIC for all possible candidates because it requires extremely computational load. Thus, we restrict the number of hyper-parameters as follows:

PMLE$_1$**:** All error variances have the same hyper-parameter λ_1.

PMLE$_2$**:** The error variances for observable variables have a hyper-parameter λ_1 and those for latent variables have a hyper-parameter λ_2.

It can be seen that PMLE$_1$ is useful when all of the variances have similar values while PMLE$_2$ can be used when the error variances for observable variables and those for latent variables are completely different.

4. Monte Carlo simulations

Monte Carlo simulations are conducted to investigate the performance of our proposed procedure. In this simulation study, latent variables are ξ, η, and observable variables are given by y_1, y_2, x_1, x_2. The true model is

$$\begin{bmatrix}\xi\\ \eta\\ y_1\\ y_2\\ x_1\\ x_2\end{bmatrix} = \begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0\\ 0.5 & 0 & 0 & 0 & 0 & 0\\ 0 & 1.0 & 0 & 0 & 0 & 0\\ 0 & 0.6 & 0 & 0 & 0 & 0\\ 0.7 & 0 & 0 & 0 & 0 & 0\\ 0.7 & 0 & 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\xi\\ \eta\\ y_1\\ y_2\\ x_1\\ x_2\end{bmatrix} + \begin{bmatrix}\varepsilon_\xi\\ \varepsilon_\eta\\ \varepsilon_{y_1}\\ \varepsilon_{y_2}\\ \varepsilon_{x_1}\\ \varepsilon_{x_2}\end{bmatrix}, \tag{7}$$

and the true variance-covariance matrix $\Sigma_{\boldsymbol{u}}$ of $\boldsymbol{u}$ is

$$\Sigma_{\boldsymbol{u}} = \mathrm{diag}(1.0, 0.6, 0.1, 0.7, 0.5, 0.5). \tag{8}$$

Table 1: Frequency of improper solutions (Freq), mean squared error for parameters (MSE), mean value of hyper-parameters ($\overline{\lambda_1}$ and $\overline{\lambda_2}$), and the mean value of the GBIC ($\overline{\text{GBIC}}$)

	Method	Freq	MSE ($\times 10$)	$\overline{\lambda_1}$ ($\times 10^3$)	$\overline{\lambda_2}$ ($\times 10^3$)	$\overline{\text{GBIC}}$ ($\times 10^{-3}$)
	MLE	39	6.688	—	—	—
$N = 100$	PMLE_1	0	2.267	3.880	—	1.077
	PMLE_2	0	2.516	4.112	3.836	1.077
	MLE	24	4.223	—	—	—
$N = 150$	PMLE_1	0	1.485	2.588	—	1.612
	PMLE_2	0	1.563	2.922	2.615	1.612
	MLE	36	2.109	—	—	—
$N = 200$	PMLE_1	0	0.869	1.928	—	2.139
	PMLE_2	0	0.879	2.420	1.808	2.139

To estimate the true model, we assume the hypothetic model

$$\begin{bmatrix} \xi \\ \eta \\ y_1 \\ y_2 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \gamma & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & a_{y_2} & 0 & 0 & 0 & 0 \\ a_{x_1} & 0 & 0 & 0 & 0 & 0 \\ a_{x_2} & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \xi \\ \eta \\ y_1 \\ y_2 \\ x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_\xi \\ \varepsilon_\eta \\ \varepsilon_{y_1} \\ \varepsilon_{y_2} \\ \varepsilon_{x_1} \\ \varepsilon_{x_2} \end{bmatrix},$$

with variance-covariance matrix

$$\Sigma_{\boldsymbol{u}} = \text{diag}(1, \sigma_\eta, \sigma_{y_1}, \sigma_{y_2}, \sigma_{x_1}, \sigma_{x_2}).$$

In this case, the parameter vector is given by $\boldsymbol{\theta} = (\gamma, a_{y_2}, a_{x_1}, a_{x_2}, \sigma_\eta, \sigma_{y_1}, \sigma_{y_2}, \sigma_{x_1}, \sigma_{x_2})'$. The true variance-covariance matrix $\Sigma = GT\Sigma_{\boldsymbol{u}}T'G'$ is calculated by using Equations (7) and (8), and then the data were generated 100 times with sample size N ($N = 100, 150, 200$).

We compare the performance of our proposed procedure with that of maximum likelihood method. Table 1 shows the frequency of improper solutions, mean squared error (MSE) for parameters, mean value of hyper-parameters ($\overline{\lambda_1}$ and $\overline{\lambda_2}$), and the mean value of the GBIC ($\overline{\text{GBIC}}$) for MLE, PMLE_1 and PMLE_2. The mean squared error (MSE) is given by

$$\text{MSE} = \frac{1}{100}\sum_{d=1}^{100} \|\hat{\boldsymbol{\theta}}^{(d)} - \boldsymbol{\theta}_0\|^2,$$

where $\boldsymbol{\theta}_0$ are true values of $\boldsymbol{\theta}$, i.e. $\boldsymbol{\theta}_0 = (0.5, 0.6, 0.7, 0.7, 0.6, 0.1, 0.7, 0.5, 0.5)'$, and $\hat{\boldsymbol{\theta}}^{(d)}$ are the estimates of parameters for d-th dataset. From Table 1, we can see that each procedure becomes better in terms of minimizing the MSE as N increase. For each

Table 2: Mean squared error (MSE) for each parameter of coefficients

	Method	$\gamma(\times 10^3)$	$a_{y_2}(\times 10^2)$	$a_{x_1}(\times 10^2)$	$a_{x_2}(\times 10^2)$
	MLE	8.738	5.479	1.435	1.255
$N = 100$	PMLE_1	8.813	5.011	1.320	1.138
	PMLE_2	9.555	5.985	1.381	1.178
	MLE	8.182	2.796	0.849	0.871
$N = 150$	PMLE_1	8.671	2.091	0.905	0.936
	PMLE_2	8.873	2.192	0.941	0.964
	MLE	4.649	1.848	0.446	0.682
$N = 200$	PMLE_1	4.666	1.005	0.503	0.712
	PMLE_2	4.688	0.960	0.533	0.756

Table 3: Mean squared error (MSE) for each parameter of error variances

	Method	$\sigma_\eta(\times 10)$	$\sigma_{y_1}(\times 10)$	$\sigma_{y_2}(\times 10^2)$	$\sigma_{x_1}(\times 10^2)$	$\sigma_{x_2}(\times 10^2)$
	MLE	2.545	2.463	2.047	3.182	2.523
$N = 100$	PMLE_1	0.292	0.411	2.465	2.668	2.154
	PMLE_2	0.361	0.459	2.418	2.806	2.243
	MLE	1.691	1.551	1.069	1.662	1.754
$N = 150$	PMLE_1	0.257	0.285	1.152	1.677	1.807
	PMLE_2	0.288	0.298	1.147	1.755	1.872
	MLE	0.780	0.686	0.708	0.861	1.409
$N = 200$	PMLE_1	0.135	0.154	0.739	0.940	1.435
	PMLE_2	0.139	0.144	0.703	0.997	1.540

N, we obtained improper solutions several times for maximum likelihood procedure, whereas our proposed method prevented the occurrence of improper solutions for all datasets. Moreover, the MSE of PMLE_1 is much smaller than that of MLE, which means the Bayesian approach yields more stable estimates than maximum likelihood technique. In addition, the values of $\overline{\lambda_1}$, $\overline{\lambda_2}$ for PMLE_2 are almost the same, and they are also similar to the value of $\overline{\lambda_1}$ for PMLE_1. Thus, it seems that PMLE_1 and PMLE_2 selected almost the same models.

Tables 2 and 3 show the mean squared error (MSE) for each parameter. When we compared the MSE of PMLE_1 with that of MLE, a large difference occurred in σ_η and σ_{y_1}. Regarding the σ_{y_1}, we obtained improper solutions several times since the true value of σ_{y_1} is relatively small compared with other parameters of error variances. Consequently, the maximum likelihood estimate of σ_{y_1} is very unstable. On the other hand, the proposed procedures PMLE_1 or PMLE_2 produced more stable estimates than MLE. For error variance σ_η, the proposed methods also provide much better estimates

Table 4: The result of GFI, AGFI and GBIC and corresponding hyper-parameters λ_1 and λ_2 for MLE, PMLE_1 and PMLE_2 for the hypothetic model given in Figure 2.

	MLE	PMLE_1	PMLE_2
GFI	0.9037	0.9038	0.9038
AGFI	0.8736	0.8738	0.8738
λ_1	—	0.0014	0.0006
λ_2	—	—	0.0017
GBIC	—	23494	23492

than MLE.

As a result, our proposed method prevents the occurrence of improper solutions and also yields stable estimates. Also, the result of PMLE_1 is very similar to that of PMLE_2. This is because the observable variables and exogenous variables are usually normalized and then they may not have completely different error variances.

5. Real data example

We applied the proposed modeling procedure to the slump of personal consumption dataset (Toyoda, 1998). This data set was surveyed from 10/06/1998 to 15/07/1998. During this long period, Japan was suffering a slump. The aim of this analysis is to find out the cause of the recession by conducting a causality analysis. This dataset consists of $N = 405$ samples and $p = 21$ observable variables $X_1, \cdots, X_{21}$, and Toyoda (1998) considered 8 latent variables $F_1, \cdots, F_8$. The 8 latent variables and corresponding observable variables are given in Appendix B.

We made a hypothetic model based on Toyoda (1998), which is given in Figure 2. First, the model of Figure 2 is estimated by maximum likelihood procedure, which is given in Figure 3. This procedure produced improper solutions since the error variance for X_{15} was -0.046. The standard error of error variance for X_{15} was 0.649, and thus the 95% confidence interval includes 0. This means the cause of improper solutions might be sampling fluctuation (see, e.g., van Driel, 1978; Chen *et al.*, 2001).

In order to prevent the occurrence of improper solutions, we applied the proposed procedures (PMLE_1 and PMLE_2) to this dataset. The estimates in PMLE_1 and PMLE_2 were respectively given in Figure 4 and Figure 5. The result of GFI, AGFI and GBIC and corresponding hyper-parameters λ_1 and λ_2 for each procedure is also given in Table 4.

From Figure 4, the penalized maximum likelihood estimate of error variance for X_{15} was positive, which means the proposed procedure prevented the occurrence of improper solutions. Additionally, Figure 4 and 5 indicate the estimates in PMLE_2 and PMLE_1 are very similar. Moreover, the result of GFI and AGFI are also almost the same. This means it is not necessary to assume that error variances for observable variables and those for latent variables have different hyper-parameters.

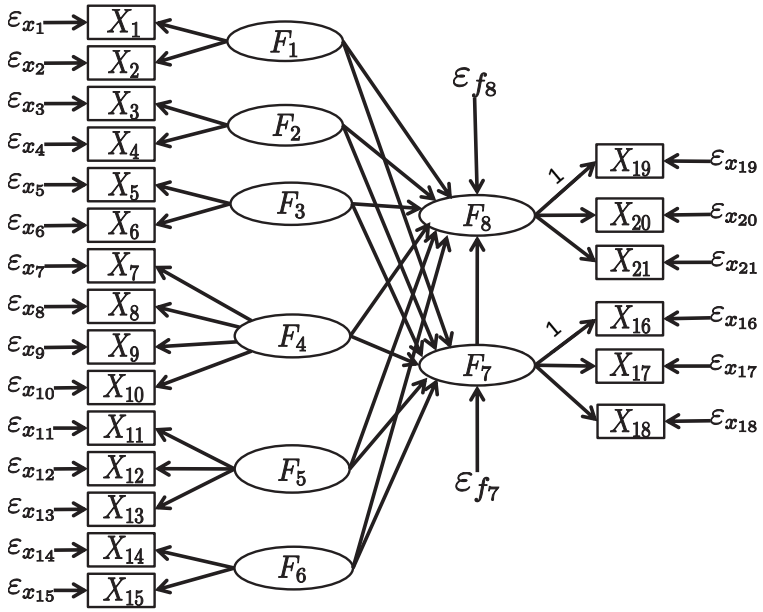

Figure 2: Hypothetic model for slump of personal consumption data

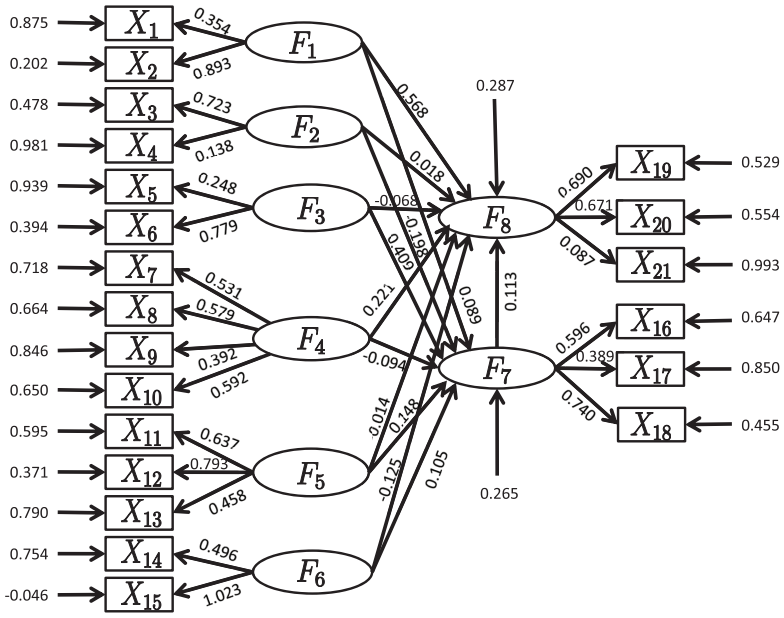

Figure 3: Maximum likelihood estimates

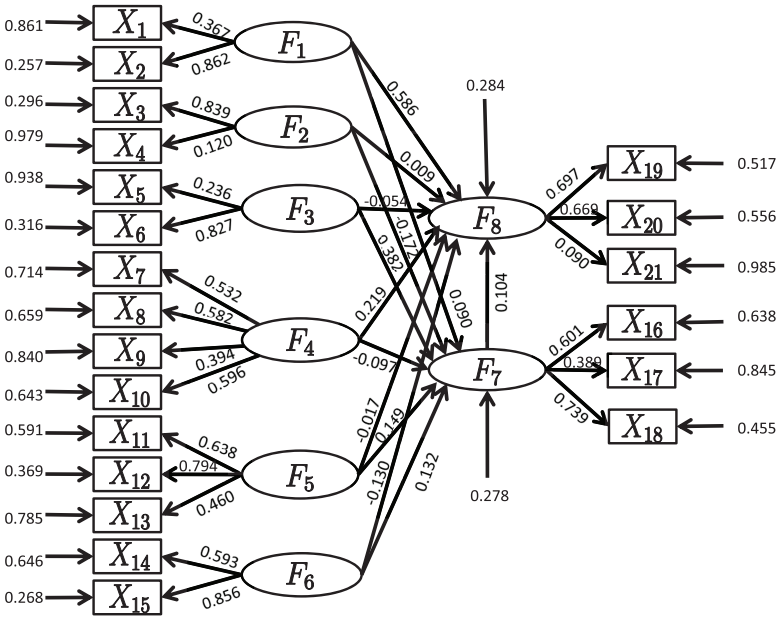

Figure 4: PMLE_1

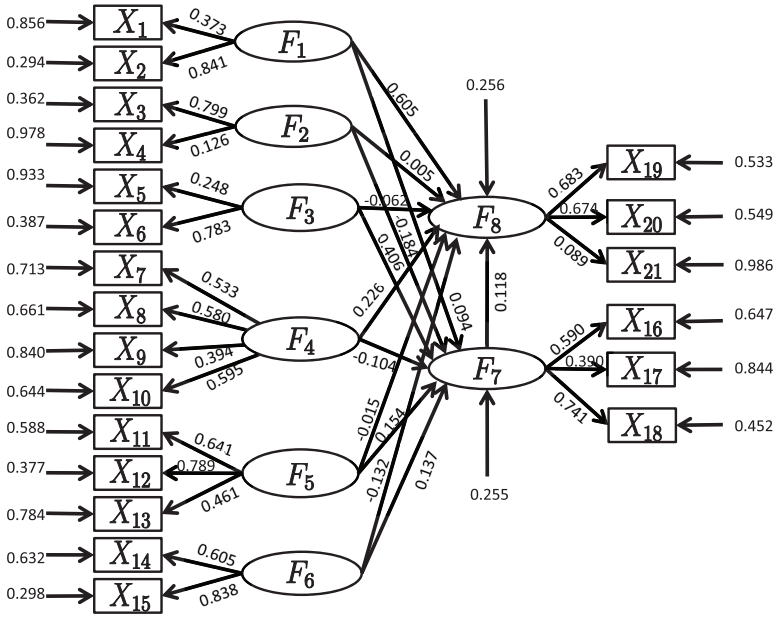

Figure 5: PMLE_2

6. Concluding and remarks

In the structural equation modeling, the maximum likelihood estimates of error variances can turn out to be zero or negative. In order to overcome this difficulty, the Bayesian approach is employed by specifying a prior distribution for error variances. Crucial issues in this modeling procedure include the choice of hyper-parameters in the prior distribution. We derived a model selection criterion from a Bayesian point of view to select these parameters. In order to reduce a computational load of the proposed modeling procedure, we considered two kinds of restrictions: all error variances have the same hyper-parameter (PMLE_1), and the error variances for observable variables and those for latent variables have different hyper-parameters (PMLE_2). The proposed procedure was applied to artificial data, and we found that our method prevents the occurrence of improper solutions and provides stable estimates. Our modeling strategy is applied to the slump of personal consumption data, and also prevented the occurrence of improper solutions. For both artificial and real data, PMLE_1 and PMLE_2 yield almost the same result. Therefore, from a practical point of view, the PMLE_1 may be preferable since it does not need computation time compared with PMLE_2.

The structural equation modeling is usually used to investigate the linear relationships among observed variables and latent variable. However, models that have nonlinear structure are often encountered in social and behavioral sciences (see, e.g., Lee and Zhu, 2003). As a future research topic, it is interesting to propose a selection procedure of hyper-parameters for nonlinear structural equation modeling. Another important topic is to select not only the value of hyper-parameter, but also the structural equation model itself. The lasso (Tibshirani, 1996) is one way to achieve this, since it can produce some coefficients that are exactly zero and then the corresponding paths are automatically eliminated. It is of our interest to apply the lasso and its related regularization methods to the structural equation models.

Appendix A: Derivation of $J_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$

This appendix derives the $J_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ included in the second differential of penalized log-likelihood function. First, we define a function F of $\boldsymbol{\theta}$:

$$F = \log|\Sigma| + \mathrm{tr}(\Sigma^{-1}S).$$

The relationship between F and second differential of log-likelihood function in (2) can be obtained as follows:

$$\frac{\partial^2 \log f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} = -\frac{N}{2}\frac{\partial^2 F}{\partial\theta_i\partial\theta_j} \quad (i, j = 1, \cdots, k).$$

Hence, if we derive a second differential of F, the second differential of log-likelihood function can be obtained. It is known that the second differential of F is given by (Lee and Jennrich, 1979):

$$\frac{\partial^2 F}{\partial\theta_i\partial\theta_j} = \mathrm{tr}\Sigma^{-1}\dot{\Sigma}_i\Sigma^{-1}\dot{\Sigma}_j + \mathrm{tr}\Sigma^{-1}(\Sigma - S)\Sigma^{-1}(\ddot{\Sigma}_{ij} - 2\dot{\Sigma}_i\Sigma^{-1}\dot{\Sigma}_j),$$

where

$$\dot{\Sigma}_j = \frac{\partial \Sigma}{\partial \theta_j}, \tag{A1}$$

$$\ddot{\Sigma}_{ij} = \frac{\partial^2 \Sigma}{\partial \theta_i \partial \theta_j}. \tag{A2}$$

For structural equation modeling, Equations (A1) and (A2) are given by

$$\begin{aligned}
\frac{\partial \Sigma}{\partial a_{\alpha\beta}} &= GT\Delta_{\alpha\beta}T\Sigma_{\boldsymbol{u}}T'G' + GT\Sigma_{\boldsymbol{u}}T'\Delta_{\beta\alpha}T'G', \\
\frac{\partial \Sigma}{\partial \sigma^u_{wx}} &= GT\Delta_{wx}T'G', \qquad \text{(A3)} \\
\frac{\partial^2 \Sigma}{\partial a_{\gamma\delta}\partial a_{\alpha\beta}} &= GT(\Delta_{\gamma\delta}T\Delta_{\alpha\beta}T\Sigma_{\boldsymbol{u}} + \Delta_{\alpha\beta}T\Delta_{\gamma\delta}T\Sigma_{\boldsymbol{u}} + \Delta_{\alpha\beta}T\Sigma_{\boldsymbol{u}}T'\Delta_{\delta\gamma})T'G' \\
&\quad + GT(\Delta_{\gamma\delta}T\Sigma_{\boldsymbol{u}}T'\Delta_{\beta\alpha} + \Sigma_{\boldsymbol{u}}T'\Delta_{\delta\gamma}T'\Delta_{\beta\alpha} + \Sigma_{\boldsymbol{u}}T'\Delta_{\beta\alpha}T'\Delta_{\delta\gamma})T'G', \\
\frac{\partial^2 \Sigma}{\partial \sigma^u_{wx}\partial a_{\alpha\beta}} &= GT(\Delta_{\alpha\beta}T\Delta_{wx} + \Delta_{wx}T'\Delta_{\beta\alpha})T'G', \qquad \text{(A4)} \\
\frac{\partial^2 \Sigma}{\partial \sigma^u_{yz}\partial \sigma^u_{wx}} &= \boldsymbol{O}_{p\times p},
\end{aligned}$$

where Δ_{ij} is a matrix with one on (i,j)-th element and zeros elsewhere, and σ^u_{xy} is the (x,y)-th element of $\Sigma_{\boldsymbol{u}}$. Note that σ^u_{xy} corresponds to σ^u_i. For example, when the error variance σ^u_1 is the $(2,3)$-th element of $\Sigma_{\boldsymbol{u}}$, $\sigma^u_1 = \sigma^u_{23}$. The reason why we use the notation σ^u_{xy} is that it is needed to derive the Equations (A3) and (A4).

To derive $J_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$, we need to obtain the second differential of the logarithm of prior distribution $\log \pi(\boldsymbol{\theta})$ regarding error variances, which is given by

$$\frac{\partial^2 \log \pi(\boldsymbol{\theta})}{\partial (\sigma^u_i)^2} = \frac{2}{(\sigma^u_i)^2} - \frac{2N\lambda_i}{(\sigma^u_i)^3}, \quad (i = 1, \cdots, v). \tag{A5}$$

The second differential for other parameters is zero. Then, the (i,j)-th element of $J_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ in (6) is given by

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \frac{1}{2}\frac{\partial^2 F}{\partial \theta_i \partial \theta_j} - \frac{1}{N}\frac{\partial^2 \log \pi(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\partial \theta_i \partial \theta_j}. \tag{A6}$$

The first term is calculated by using (A2), and the second term is given by (A5).

Appendix B: Description of real data

The 8 latent variables and corresponding observable variables for the slump of personal consumption dataset are given in the following:

- F_1: Changes in income
  - — X_1: Increase and decrease in income
  - — X_2: Consciousness of joy and sorrow for life

- F_2: Fears of a recession
  — X_3: Sense for the state of the economy
  — X_4: Prospect of the state of the economy

- F_3: Expectation for decrease in price
  — X_5: Prospect of decrease in price
  — X_6: The number of products that have a prospect of decrease in price

- F_4: Saturation of consumption
  — X_7: Getting away from from shopping
  — X_8: Overmuch fullness
  — X_9: Getting tired of shopping
  — X_{10}: Lack of fascinating products

- F_5: Prospect of future of society
  — X_{11}: Society of guarantee for position
  — X_{12}: Society of increase and decrease in income
  — X_{13}: Society of guarantee for old people

- F_6: Self-searching for life
  — X_{14}: Self-searching for luxury
  — X_{15}: Self-searching for shopping

- F_7: Buyer motivate
  — X_{16}: The number of goods that have appetites
  — X_{17}: A great desire to buy
  — X_{18}: The number of goods that have appetites if the price is down

- F_8: Buyer behavior
  — X_{19}: Limits on spending
  — X_{20}: The number of goods that limit on spending
  — X_{21}: The rate of realization for purchasing

For detail of the explanation of each variables, we refer to Toyoda (1998).

Acknowledgment

The authors would like to thank the anonymous reviewer for the helpful comments and suggestions.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, **52**, 317–332.

Anderson, J. C. and Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, **49**, 155–173.

Bentler, P. M. and Stein, J. A. (1992). Structural equation models in medical research. *Stat. Methods Med. Res.*, **1**, 159–181.

Bentler, P. M. and Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika*, **45**, 289–308.

Bock, R. D. and Bargmann, R. E. (1966). Analysis of covariance structures. *Psychometrika*, **31**, 507–534.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. *Psychometrika*, **50**, 229–242.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J. and Kirby, J. B. (2001). Improper solutions in structural equation models, causes, consequences, and strategies. *Soc. Methods Res.*, **29**, 468–508.

Flora, D. B. and Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol. Methods*, **9**, 466–491.

Gerbing, D. W. and Anderson J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika*, **52**, 99–111.

Hirose, K., Kawano, S., Konishi, S. and Ichikawa, M. (2010). Bayesian information criterion and selection of the number of factors in factor analysis models. *To appear in Journal of Data Science.*

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, **57**, 239–251.

Jöreskog, K. G. and Sörbom, D. (1996). *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Scientific Software International: Hove and London.

Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, **91**, 27–43.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling.* New York: Springer.

Lee, S. Y. and Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika*, **44**, 99–113.

Lee, S. Y. and Zhu, H. T. (2003). Statistical analysis of nonlinear structural equation models with continuous and polytomous data. *Br. J. Math. Stat. Psychol.*, **53**, 209–232.

Liu, B., de la Fuente, A. and Hoeschele, I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, **178**, 1763–1776.

Martin, J. K. and McDonald, R. P. (1975). Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. *Psychometrika*, **40**, 505–517.

McArdle, J. J. (1980). Causal modeling applied to psychonomic systems simulation. *Behav. Res. Methods Instrum.*, **12**, 193–209.

McArdle, J. J. and McDonald, R. P. (1984). Some algebraic properties of the reticular

action model for moment structures. *Br. J. Math. Stat. Psychol.*, **37**, 234–251.

Ninomiya, Y., Yanagihara, H. and Yuan, K.-H. (2008). Selecting the number of factors in exploratory factor analysis via locally conic parameterization. *Research Memorandum No. 1078, The Institute of Statistical Mathematics*, Tokyo.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, **58**, 267–288.

Pugesek, B. H., Tomer, A. and von Eye, A. (2003). *Structural Equation Modeling Applications in Ecological and Evolutionary Biology.* New York: Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Toyoda, H. (1998). *Covariance Structure Analysis [Case Examples] — Structural Equation Modeling— (in Japanese).* Kitaohji-shobo Publishing Co., Ltd.

van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika*, **43**, 225–243.

Xiong, M., Li, J. and Fang, X. (2004). Identification of genetic networks. *Genetics*, **166**, 1037–1052.

Yoshida, R and West, M. (2010). Bayesian learning in sparse graphical factor models via annealed entropy. *J. Mach. Learn. Res.*, **11**, 1771–1798.

Received May 24, 2010
Revised November 24, 2010